Two-Stage Approach for Semantic Image Segmentation of Breast Cancer : Deep Learning and Mass Detection in Mammographic images

Faycal TOUAZI[a], Djamel GACEB[a], Marouane CHIRANE[a], Selma HERZALLAH[a]

[a] *LIMOSE laboratory, Computer science department, University M'hamed Bougara, Independence Avenue, 35000 Boumerdes, Algeria*

Abstract

Breast cancer is a significant global health problem that predominantly affects women and requires effective screening methods. Mammography, the primary screening approach, presents challenges such as radiologist workload and associated costs. Recent advances in deep learning hold promise for improving breast cancer diagnosis. This paper focuses on early breast cancer detection using deep learning to assist radiologists, reduce their workload and costs. We employed the CBIS-DDSM dataset and various CNN models, including YOLO versions V5, V7, and V8 for mass detection, and transformer-based (nested) models inspired by ViT for mass segmentation. Our diverse approach aims to address the complexity of breast cancer detection and segmentation from medical images.

Our results show promise, with a 59% mAP50 for cancer mass detection and an impressive 90.15% Dice coefficient for semantic segmentation. These findings highlight the potential of deep learning to enhance breast cancer diagnosis, paving the way for more efficient and accurate early detection methods.

Keywords 1

Breast Cancer, Deep Learning, ViT, NEST, YOLO

1. Introduction

Breast cancer remains one of the most prevalent diseases among women globally and stands as a leading cause of mortality in gynecological cancers. Across the world, the situation is indeed dire, with one in ten women affected by this disease during their lifetime. It ranks second in overall cancer incidence, following prostate cancer, affecting individuals of all genders. Despite considerable efforts in the form of screening programs aimed at prevention and early detection, there is an urgent need to enhance methods for analyzing mammography images.

Mammography represents the unquestionable gold standard for breast exploration, offering unmatched performance in breast cancer surveillance and early detection. Each year, millions of mammograms are produced worldwide for the early screening of breast cancer or to establish a diagnosis to guide therapeutic interventions. However, the interpretation of these images remains a major challenge for healthcare professionals, as they provide complex radiological information that is challenging to fully exploit through human expertise, which relies on visual interpretation and experience.

Confronted with this challenge, the development of dedicated software for mammography image analysis becomes imperative to optimize their utilization for the benefit of both patients and physicians. A more suitable method of interpretation is required to enable earlier detection and more effective management of the disease.

IDDM'2023: 6th International Conference on Informatics & Data-Driven Medicine, November 17 - 19, 2023, Bratislava, Slovakia
EMAIL: f.touazi@univ-boumerdez.dz (A. 1); d.gaceb@univ-boumerdez.dz (A. 2), ch.marouanee@gmail.com (A. 3), harzallahselma@gmail.com (A. 4)
ORCID: 0000-0001-5949-5421 (A. 1);

Deep Learning (DL) has revolutionized various real-world domains by providing accurate and powerful solutions. In the medical field, it also offers promising solutions for the interpretation of medical images, allowing for highly precise analysis. This paper project focuses on applying Deep Learning using a range of models and techniques, including transformers, with the goal of detecting breast cancer in mammograms. To achieve this objective, we integrate the YOLO (You Only Look Once) model [1] for precise detection of regions of interest (ROI) in mammographic images. Once the regions of interest are identified, we employ SegNest, an adaptation of the ViT Nest model [2] for semantic segmentation to perform semantic segmentation of tumors.

In this paper, our primary objective is to push the boundaries of early breast cancer detection by harnessing the advancements in Deep Learning and computer vision. To achieve this goal, we will apply techniques of object detection and semantic segmentation to effectively localize and characterize mass breast cancer in real mammography images. Moreover, we will explore a hybrid approach that combines transformers (ViT) with CNN to leverage their respective strengths in breast cancer detection.

The paper is structured as follows: In Section 2, we delve into the related work in the field, providing a thorough review of existing literature to establish the context and significance of our research. Section 3 outlines our proposed approach, elucidating the methodology and techniques employed in our study. The heart of our contribution lies in Section 4, where we present our results and engage in an in-depth discussion, offering insights and interpretations of the data. Finally, in Section 5, we draw our conclusions, summarizing the key findings, their implications, and potential avenues for future research.

2. Related works

In this section, we present an overview of recent studies in the field of breast cancer detection and tumor segmentation using deep learning techniques.

For the breast cancer detection, the authors of [3] proposed a two-step method using high-resolution mammograms. They achieved a significant improvement over Faster R-CNN in terms of detection accuracy for BI-RADS categories. Hamed Aly et al. [4] applied YOLO-V3 for automated breast mass detection, achieving a mass detection rate of 89.4\% and high precision for classifying malignant and benign masses. Prinzi et al. [5] presents an approach to automated breast cancer detection using YoloV5 architecture, which reached an mAP50 of 49.8% on CBIS-DDSM dataset.

For Breast Tumor Segmentation, Soltani et al. [6] employed Mask R-CNN, reporting a promising performance with high precision of 0.75%, recall of 0.80%, and F1 score of 0.825%. Yu et al. [7] introduced Dense-Mask R-CNN, which surpassed the original Mask R-CNN in breast mass detection on the CBIS-DDSM dataset, with an average precision (AP) of 0.65.

Among the approaches based on transformers we cite the work of Liu et al. [8] introduced TrEnD, an encoder-decoder model based on transformers for mammography mass segmentation. They applied superpixel-based adaptive patch embedding and achieved improved Dice and Intersection over Union (IoU) scores on the CBIS-DDSM and INBreast datasets. Su et al. [9] developed a YOLO-LOGO model for breast mass detection and segmentation in digital mammograms. Their model effectively combined mass detection and segmentation using YOLOV5L6 and a Vision Transformer (ViT), showing promising results that outperformed other segmentation models. They trained their model on CBIS-DDSM dataset and they achieved a dice score of 84.49%.

Prezi et al . [5] proposed an approach for breast cancer detection in CBIS-DDSM mammograms. The study compares various YOLO architectures namely YOLO V3 YOLO V5 and YOLOV5-Transformer. Within this architecture, the Transformer block was incorporated into the second-to-last layer of the backbone network, specifically positioned among the trio of convolutional layers that precede the spatial pyramid pooling layer. The small YOLOV5 model outperforms others with a mAP of 0.621.

As summary of this related works, the hybridization of YOLO with Vision Transformer (ViT) represents a promising avenue for breast cancer detection and tumor segmentation, as evidenced by the compelling results obtained in the existing literature. This fusion of YOLO and ViT architectures has consistently demonstrated superior performance in various studies, underscoring its potential to enhance both mass detection and segmentation tasks.

3. Proposed approach

In this section, we describe our proposed approach of the detection and diagnosis of breast cancer, based on deep learning. The proposed approach focuses specifically on the detection of breast masses within the context of breast cancer. It is a holistic approach that combines mass detection stage using YOLO architecture and segmentation stage using SegNesT architecture. By integrating these two stages (see Figure 1), this approach seamlessly integrates different aspects of deep learning to create a more holistic and potentially more effective diagnostic system for patients.

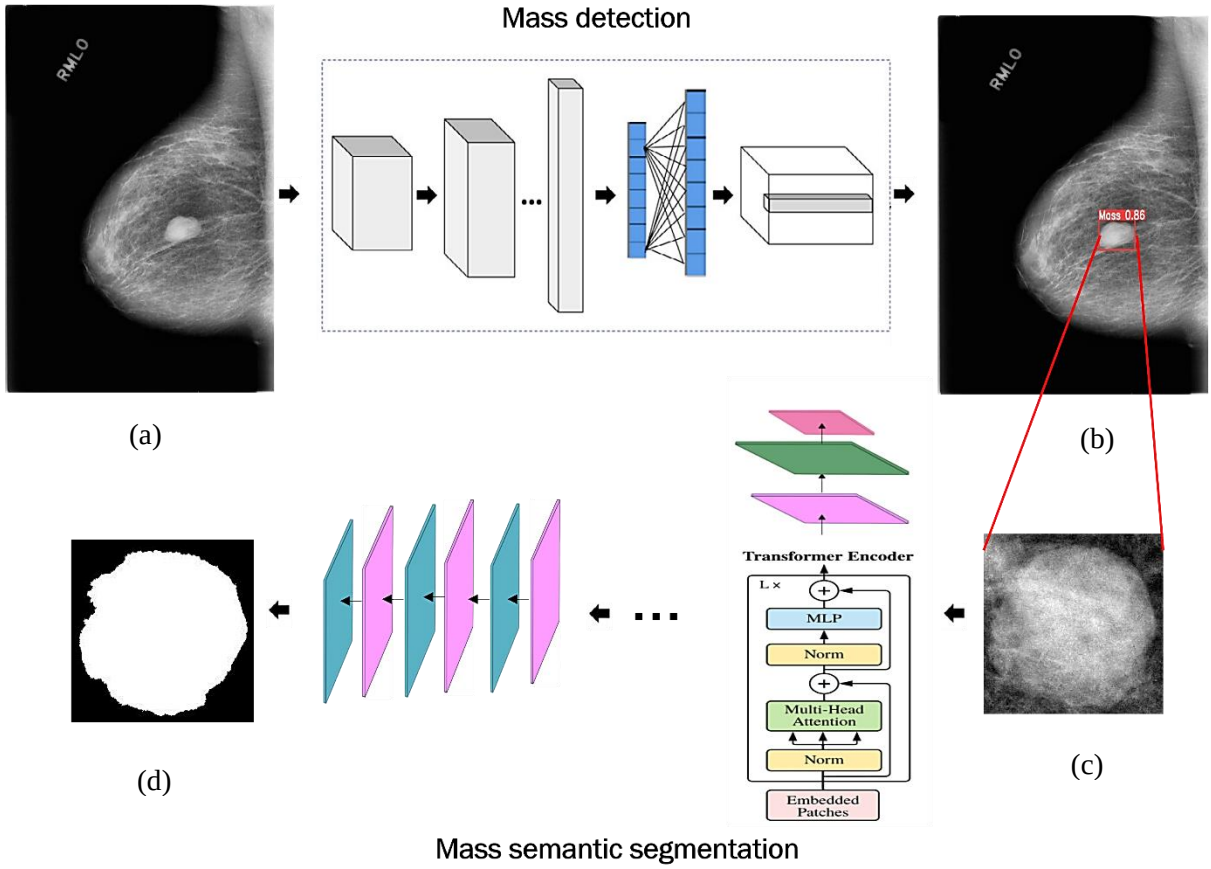


Figure 1: The framework of the proposed approach. (a) Original mammogram of the mass, (b) Detected Region of Interest (ROI) of the mass, (c) ROI of the detected mass, (d) Binary mask segmented from the ROI of the mass

3.1. Breast mass cancer detection based on YOLO model

At this level, a comprehensive comparative study of common object detection methods is conducted. Among the various approaches examined, YOLO [1] emerged as a promising choice due to its advanced real-time object detection performance. In the first phase of proposed approach for breast cancer detection (detected region of interest ROI of the mass), three most recent versions of the YOLO architecture V5 [10], V7 [11] and V8 [12] are used and compared (see Figures 2, 3 and 4 for architecture details).

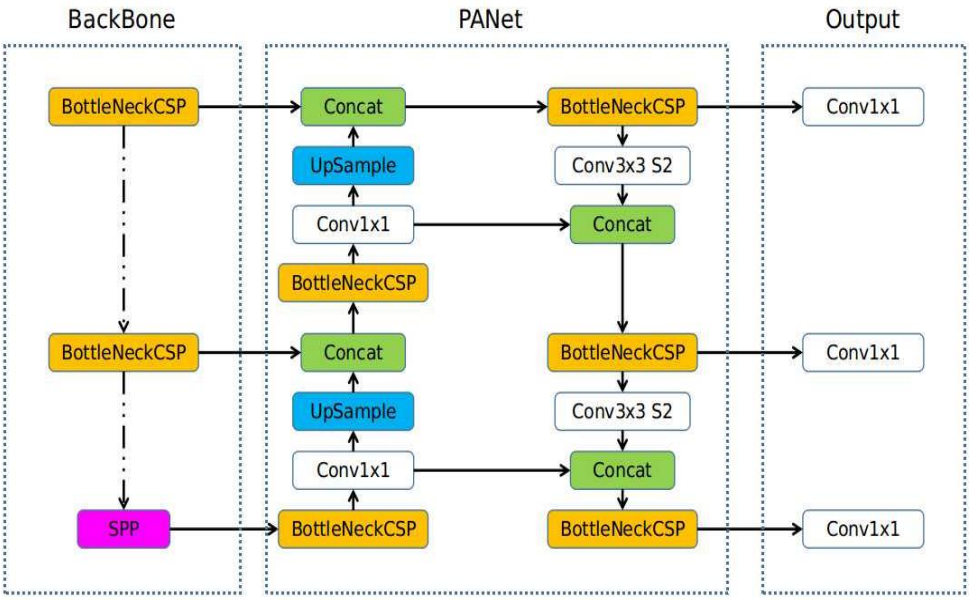


Figure 2: YOLO V5 architecture [13]

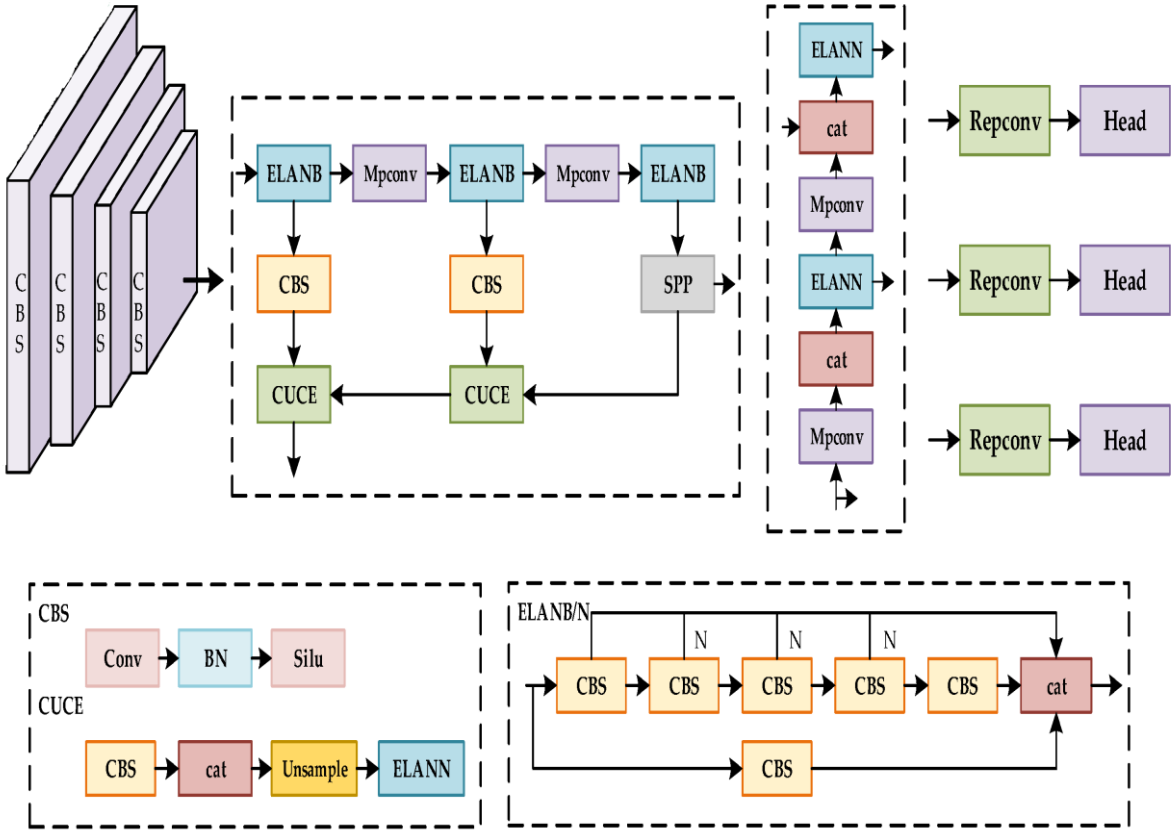


Figure 3: YOLO V7 architecture [14]

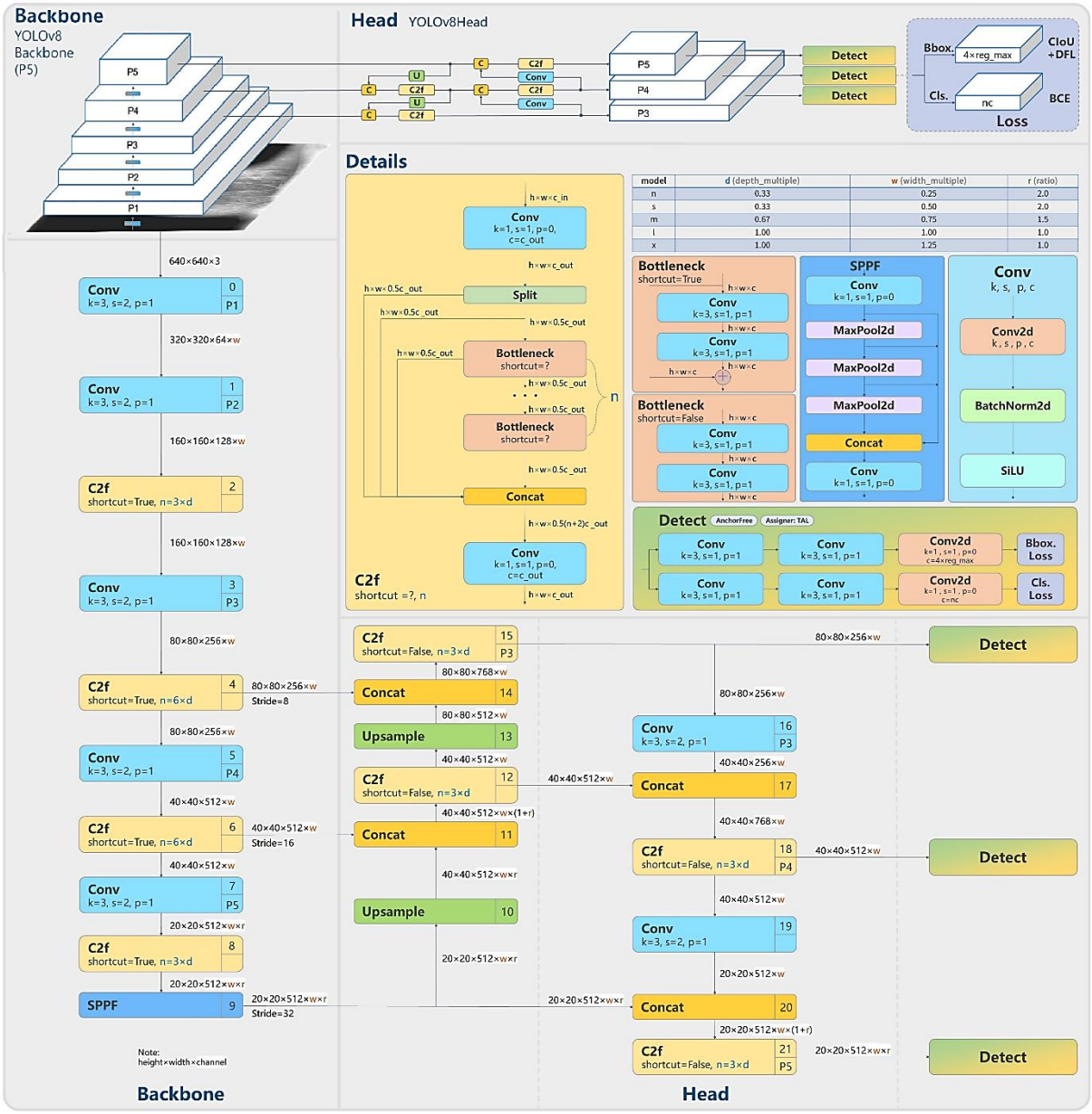


Figure 4: YOLO V8 architecture [15]

3.2. Breast cancer segmentation based on SegNesT architecture

Once these regions of interest have been identified (in the first stage), we applied, in the second stage, image segmentation using the SegNesT model. This model is a customized version of the ViT NEST [2], adapted specifically for effective segmentation tasks. SegNesT excels in precisely outlining the contours of relevant structures, thereby enhancing lesion characterization.

This architecture adopts a hierarchical approach based on the Transformer architecture for image processing. The workflow start from data pre-processing, where an image and its corresponding mask (label) are fed into the model. The image is initially partitioned into patches, which facilitates the capture of local details while retaining a global image representation and accommodating different resolutions. Subsequently, the model employs multiple hierarchical NesT levels to capture information across various scales. Each hierarchical level comprises a pooling layer, a convolutional with normalization layer, a position embedding layer and a transformer layer to model feature dependencies. Ultimately, this model can represent intricate information at multiple resolutions. Finally, it employs a

deprojection operation (un-patchify) to reconstruct the image (mask) based on the extracted features (see Figure 5).

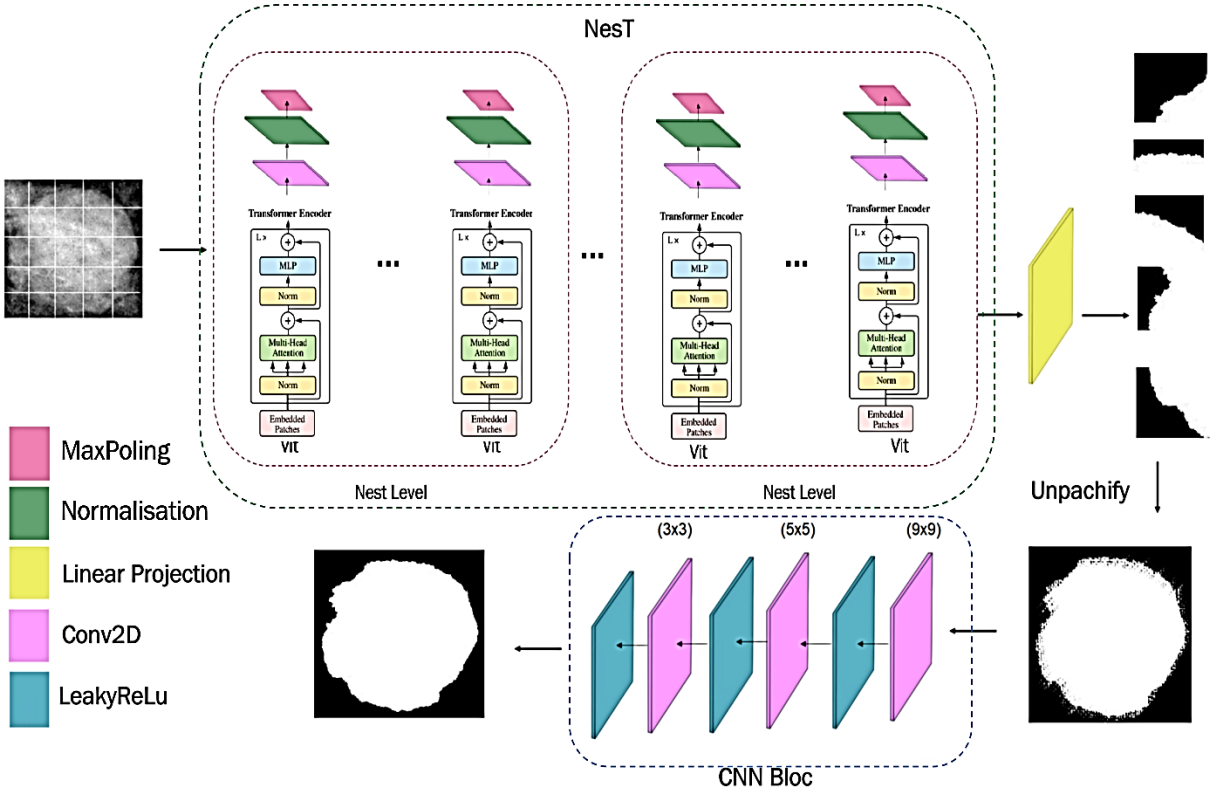


Figure 5 : SegNest architecture

Our architectural design encompasses three primary components:

- **NesT (Nested Transformers)**: The NEST VIT architecture encompasses five crucial components for comprehensive image analysis. Firstly, it initiates with Patch Embedding, dividing the input image into smaller patches and transforming them into embeddings, facilitating the processing of both local and global information. The architecture then operates across multiple Hierarchical Levels, focusing on feature extraction at various scales, utilizing self-attention mechanisms and feed-forward networks to enhance feature representations. To maintain spatial awareness, Positional Embeddings are incorporated and added to patch representations. Following feature extraction, a Feature Refinement stage refines feature maps using convolutional layers, effectively eliminating artifacts and enhancing visual quality through the following steps:
    - Linear Layer
    - Unpatchify Layer
    - Convolution Layer 1: Kernel size = 9x9
    - Convolution Layer 2: Kernel size = 5x5
    - Convolution Layer 3: Kernel size = 3x3
    - MaxPooling Layer: Kernel size = 3x3

  Finally, an Image Reconstruction stage rearranges feature representations into the original image format, ensuring a coherent and visually appealing final output.

- **Mask Reconstruction:** The second component focuses on the reconstruction of the image itself. It takes the embedding vectors generated by the "NesT" part and arranges them in a grid of patches to reconstruct the segmented image or mask.
- **CNN Block:** The third and essential component comprises a CNN block that contributes significantly to the overall architecture. This block includes three convolution layers.

    - Convolution Layer 1: Kernel size = 9x9
    - LeakyReLU Layer
    - Convolution Layer 2: Kernel size = 5x5
    - LeakyReLU Layer
    - Convolution Layer 3: Kernel size = 3x3

    Padding is applied in these three layers to maintain the image size after convolution. This CNN block effectively eliminates any blocking artifacts that might be present in the reconstructed image from the NesT component, resulting in a smoother and visually appealing final output.

4. Experimentations and results

4.1. Dataset used in this work

In this work we have chosen the CBIS-DDSM dataset [16], a subset of the Digital Database for Screening Mammography (DDSM) [17], is a valuable resource for breast cancer research. It stands out due to its complexity, encompassing diverse digital mammography images of both normal and abnormal cases. These images are rich in details and annotations, making it a challenging dataset for tasks like lesion detection and classification. The dataset's complexity arises from the presence of subtle lesions, varying image qualities, and diverse lesion types. We used 1253 images for the train set and 363 for the test set.

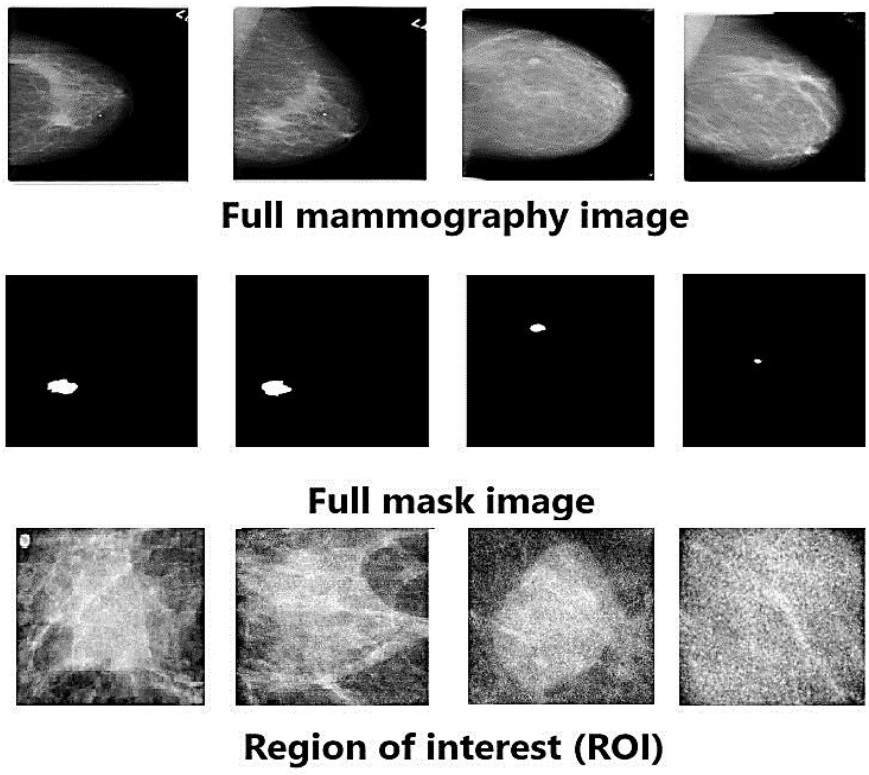


Figure 6: CBIS DDSM images examples

4.2. Data pre-processing

As part of our implementation, we applied some preprocessing techniques to prepare our dataset for use, we summarize them in the following points:

- **Image croppe**

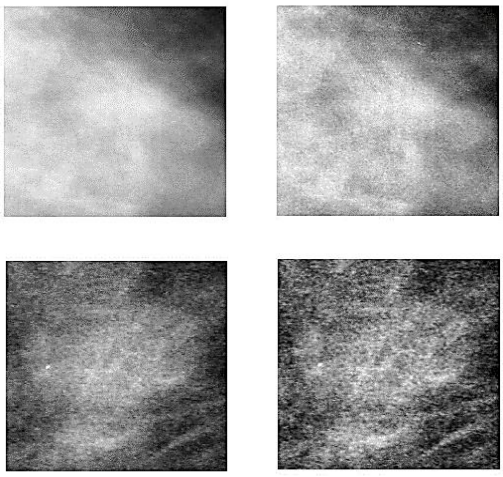
  We have applied image cropping to focus on a specific region of interest (ROI) within the image. Our approach involves utilizing the mask images supplied within the dataset, allowing us to extract and crop the white regions from the original images.

- **Resize**

  We resized all cropped images to a size of 224×224 px to fit the model input.

- **Image enhancement using CLAHE method**

  We use this technique to improves the visibility of details in an image by enhancing the contrast. It does this by redistributing the intensity values in a way that ensures a more uniform distribution of pixel values, thereby making both dark and bright regions more distinguishable.



(a) (b)

Figure 7: Examples of applying CLAHE: (a): input image, (b): CLAHE result.

- **Normalization**

  The normalization process consisted of subtracting the minimum value from each pixel then divide by the range of pixel values (maximum minus minimum). This scaling ensured that pixel values were within the desired range to improve performance, avoid numerical instabilities and allow consistent comparisons between pixel values. We applied normalization to the images input by scaling their pixel values within a normalized range of 0 to 1.

4.3. Used metrics and loss functions

Here we present the different metrics and loss functions used to train and evaluate our models.

Intersection over union (IoU)*:* is another evaluation metric used to assess the quality of segmentations. It is calculated as the ratio of the intersection area between the predicted mask and the reference mask to the union area of the two masks. The IoU is given by the formula:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Dice score (DSC): is an evaluation measure used to assess the similarity between two sets, often used to assess the quality of medical image segmentations. For two sets A and B, the Dice Score is calculated as follows:

$$DSC = \frac{2 * |A \cap B|}{|A| + |B|} \tag{2}$$

where A and B represent the cardinalities of sets A and B respectively.

Dice loss: is a loss function used for training segmentation models, especially for tasks where segmentation is represented by a binary mask. The Dice Loss is defined as the inverse of the Dice Score. The objective is to minimize this loss to improve the quality of the segmentation.

$$\mathcal{L}_{Dice} = 1 - DSC \tag{3}$$

Binary cross-entropy (BCE): is a commonly used loss function for training binary classification models. It is used when each example can belong to only one class. It measures the distance between the model's predictions and the true labels (ground truth). The binary cross-entropy is defined as follows:

$$\mathcal{L}_{BCE} = -\big(y * log(p) + (1 - y) * log(1 - p)\big) \tag{4}$$

where y is the true binary label (0 or 1) and p is the probability predicted by the model for this label.

Focal binary cross-entropy (Focal loss): is a specialized variant of the binary cross-entropy loss function. The Focal Loss was introduced to address the problem of training deep neural networks on imbalanced datasets, where the model may struggle to effectively learn from the minority class examples.

The main idea behind Focal Binary Cross-Entropy is to down-weight the loss contribution of easy-to-classify examples and focus more on the hard-to-classify ones. It does this by introducing two key parameters:γ) and α.

$$\mathcal{L}_{BCE}^{Focal} = -\alpha_{t_i}\big(1 - p_{t_i}\big)^{\gamma} \log\big(p_{t_i}\big) \tag{5}$$

Where:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{6}$$

$$\alpha_t = \begin{cases} \alpha & \text{if } y = 1 \\ 1 - \alpha & \text{otherwise} \end{cases} \tag{7}$$

Combined loss function (Combo Loss): is a composite loss function that simultaneously minimizes the Dice loss and a modified version of the cross-entropy loss. Its formula is giving as follows:

$$\mathcal{L}_{comb} = \alpha \cdot \mathcal{L}_{BCE} + (1 - \alpha) \cdot \mathcal{L}_{Dice} \tag{8}$$

We can also consider a focal version of this combined loss function as follow:

$$\mathcal{L}_{comb}^{Focal} = \alpha \cdot \mathcal{L}_{BCE}^{Focal} + (1 - \alpha) \cdot \mathcal{L}_{Dice} \tag{9}$$

Mean Average Precision (mAP) : is commonly used to analyze the performance of object detection and segmentation systems. In our work is used to evaluate mass detection models. It compares the ground-truth bounding box to the detected box. The higher the score, the more accurate the model is in its detections. It is calculated using the following formula :

$$mAP = \frac{1}{n} \sum_{k=1}^{n} AP_k \tag{10}$$

Where n is the number of classes and AP_k is the Average Precision of the classe k.

$$AP_k = \frac{TP(k)}{TP(k) + FP(k)} \tag{11}$$

with *TP(k)* is the True Positive rate of the class *k*: The model predicted a label of class *k* and matches correctly to the ground truth. *FP(k)* is the False Positive rate of the class *k*: The model predicted a label of class K, but it is not a part of the ground truth. mAP50 is Mean Average Precision calculated for the IoU threshold of *0.5*.

4.4. Results and discussion

Discussion is dedicated to the analysis and discussion of the outcomes obtained from our experiments, focusing on two main aspects: Object Detection Results and Segmentation Results.

4.4.1. Mass detection results

In this subsection, we delve into the performance of our object detection model. Table 1 presents the performance of different models: YOLOv5 Small (V5 S), YOLOv7 (V7 X), and YOLOv8 Medium (V8 M). Analyzing the mAP50 score on the test set, YOLOV8 Medium achieved the best performance with a score of 59%, surpassing YOLOv5 Small with 46% and YOLOv7X with 51%. These results suggest that YOLOV8 Medium displayed the highest performance among the three tested models.

Table 1
Mass detection results using different YOLO models

Model	map50	Image Size
YOLO V5 S	46 %	1280×1280
YOLO V7 X	51%	640×640
YOLO V8 M	**59%**	640×640

It's worth noting that despite the larger size of YOLOv5 Medium with an image resolution of 640×640 compared to YOLOv5 Small, it did not achieve satisfactory results. This can be attributed to the fact that YOLOv5s was trained on high-resolution images (1280×1280), whereas YOLOv5m was trained on lower-quality images. Unfortunately, it was not possible to evaluate the performance of YOLOv5m at a resolution of 1280×1280 due to hardware limitations. Therefore, the results of YOLOv5m at this resolution are not considered in this comparison.

Table 2 presents a comparison of various models and their performance metrics in the context of breast cancer mass segmentation.

Table 2
Comparison with related works

Paper	Model	mAP
Prinz et al. [5]	YOLO V5s	49,8%
Su et al. [9]	YOLO V5s	59%
Our	YOLO V8	59%
Su et al. [9]	YOLO V5L6	**65%**

In the study by Su et al. [9], they achieved a success rate of 59% using YOLOv5 after 1000 training epochs. In contrast, our approach also reached a 59% mAP rate, but with a different approach that required only 300 training epochs. This difference in the number of epochs suggests a relative efficiency in our approach. They achieved with YOLOv5L6 an mAP50 of 65%. However, due to hardware constraints, we were unable to use this version.

Furthermore, when compared to the article by Prinzi et al. [5], which utilized YOLOv5s with data augmentation and obtained a result of 49.8% in mAP50, our approach yielded better results.

4.4.2. **Segmentation results**

The following subsection is dedicated to the analysis of our segmentation model's performance. The results of SegNet are displayed in the table above with the different loss functions used:

Table 3
Mass segmentation results using SegNest model

Loss function	Dice
$\mathcal{L}_{BCE}$	75%
$\mathcal{L}_{comb}$	81.2%
$\mathcal{L}_{comb}^{Focal}\ (dice_{weight} = 1.0, focal_{weight} = 1.0)\ )$	89.99%
$\mathcal{L}_{comb}^{Focal}\ (dice_{weight} = 0.5, focal_{weight} = 0.5\ )$	**90.15%**

In our research, we explored different loss functions in our SegNesT model training. Initially, we have used the binary cross-entropy (BCE) loss function, our model achieved a Dice score of 75%.

However, when we adopted the combined loss function, our performance improved significantly, reaching a Dice score of 81.2%. To further enhance our model's performance, we introduced the Combined Focal Loss, using dice_weight=1.0 and focal_weight=1.0, which resulted in even better performance, with an impressive Dice score of 88.99%. Finally, by adjusting the weights to dice_weight=0.5 and focal_weight=0.5, our model achieved its best result, with a remarkable Dice score of 90.15%. These results underscore the critical importance of selecting the right loss function in improving the performance of our SegNesT model.

Table 4 presents a comparison of various models and their performance metrics in the context of breast cancer mass segmentation.

Table 4
Comparison with related works

Paper	Model	Dice
Sharif Amit Kamran et al. [34]	Swin-SFTNet	24.13%
Bouzar-Benlabiod et al. [18]	U-Net SE-RestNet-101	75%
Yuehang Wang et al. [19]	AM-MSP-cGAN	84.49%
Dongdong Liu et al. [8]	TrEnD	89.48%
Our approach	**SegNest**	**90.15%**

Our SegNesT model has achieved the higher dice score among similar literature works. Bouzar-Benlabiod et al. [18] had used Attention U-net, obtained a Dice score of 75%. Yuehang Wang et al. [19] has achieved a dice score of 84.49% using a hybrid approach combining YOLO and ViT. Dongdong Liu et al. [8] also with a ViT-based model named TrEnD achieved a Dice score of 89.48%

was achieved. It is clear that our SegNesT model has outperformed the results of related works even transformer-based models, achieving the highest Dice score of 90.15%. This superior performance highlights the effectiveness of our approach compared to previous methods in the field of medical image segmentation.

Our NesT based approach plays a crucial role in addressing the quadratic complexity issue of full self-attention in vision transformers. By introducing a hierarchical nested structure and incorporating block aggregation, NesT effectively improves data efficiency and accuracy compared to previous methods within the realm of ViT-based approaches. This progress positions our approach favorably compared to other ViT based approaches.

The block aggregation mechanism plays a central role in promoting effective inter-block communication, thereby diminishing the necessity for full self-attention at each layer. This simplification of the architectural design not only amplifies the effectiveness of training with smaller datasets but also demonstrates its utility as model size scales up, illustrating NesT's enhanced efficiency in handling larger models.

4.4.3. Result samples

In this section, we provide a comprehensive showcase of result samples obtained from our study. These samples serve as illustrative examples of the outcomes generated by our research.

Figure 8 showcases the qualitative results of our model's detection task, offering a visual representation of its performance in identifying and localizing objects of interest within the dataset. These results provide valuable insights into the accuracy and precision of our model's detection capabilities, contributing to a comprehensive assessment of its overall effectiveness.

Figure 9 presents the qualitative results of our model's segmentation task, underlining the remarkable resemblance between the ground truth mask and the predicted mask generated by our SegNest model. This compelling similarity confirm the precision and fidelity of our model's segmentation capabilities

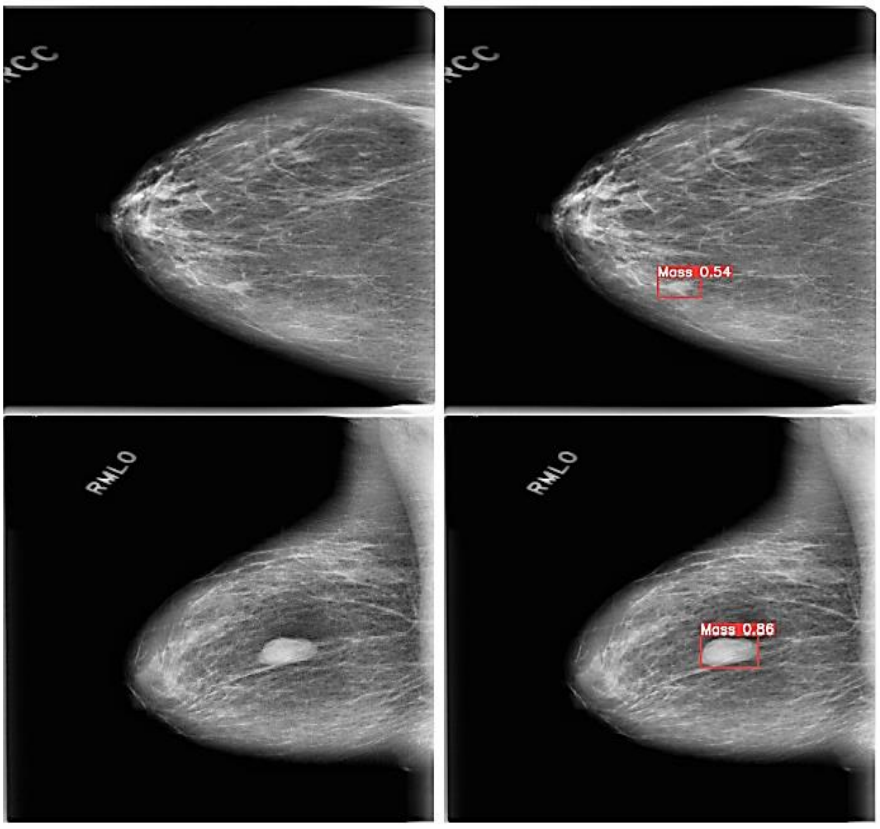
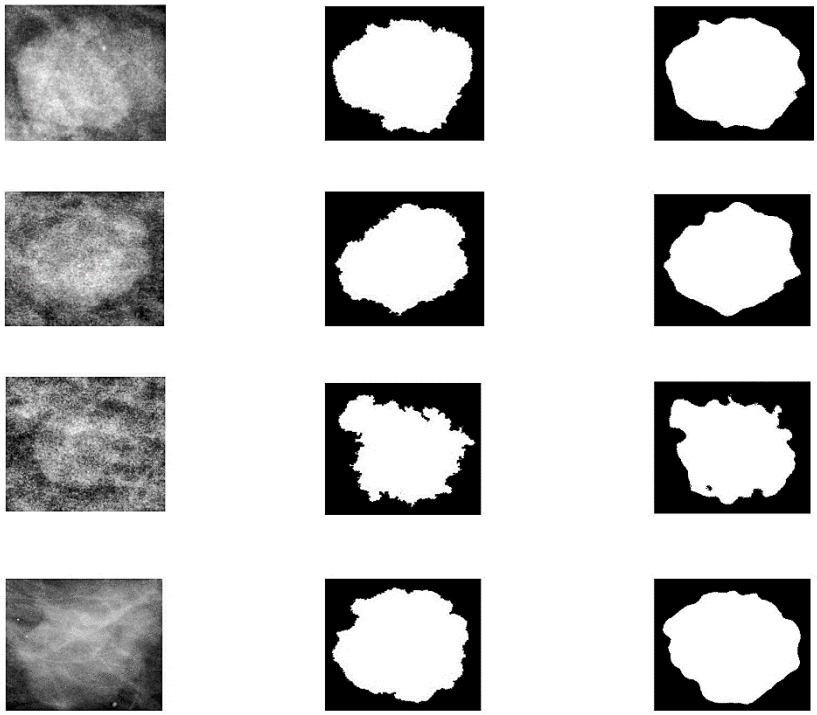


Figure 8: Masses detected with YOLOV8



(a) (b) (c)

Figure 9: Example of segmentation result with SegNest. (a) Regions of Interest (ROI), (b) Ground Truth, (c) Predicted Mask

5. Conclusion

In this paper, we highlight the potential use of object detection and semantic segmentation in the field of breast cancer detection and diagnosis. We have explored various aspects of deep learning, including mass detection using YOLO versions 5, 7, and 8, as well as breast mass cancer segmentation using our proposed SegNest architecture, based on ViT Nest. The findings indicate the efficacy of these methodologies, as YOLO V8 M achieved the highest mean average precision (mAP) of 59% among the YOLO models for mass detection. Additionally, our SegNest model demonstrated outstanding performance in mass semantic segmentation, achieving a Dice loss of 90.15%. These approaches have demonstrated their effectiveness in identifying anomalies and tumors in mammographic images, offering promising avenues for improving the accuracy of breast cancer diagnoses.

While our findings are promising, it's important to outline that our experimentation was conducted with a limited dataset. To enhance the performance and generalizability of our models, we foresee numerous directions for future research and development. These include expanding our dataset to encompass a more diverse range of cases, refining model architectures, and exploring transfer learning techniques from other medical imaging domains. These steps will be crucial in ensuring that the benefits of deep learning in breast cancer detection can be realized more broadly, ultimately benefiting both patients and healthcare professionals.

6. References

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[2] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. Ö. Arik and T. Pfister, "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[3] B. Ibrokhimov and J.-Y. Kang, "Two-stage deep learning method for breast cancer detection using high-resolution mammogram images," *Applied Sciences,* vol. 12, p. 4616, 2022.

[4] G. H. Aly, M. Marey, S. A. El-Sayed and M. F. Tolba, "YOLO based breast masses detection and classification in full-field digital mammograms," *Computer methods and programs in biomedicine,* vol. 200, p. 105823, 2021.

[5] F. Prinzi, M. Insalaco, A. Orlando, S. Gaglio and S. Vitabile, "A Yolo-Based Model for Breast Cancer Detection in Mammograms," *Cognitive Computation,* p. 1–14, 2023.

[6] H. Soltani, M. Amroune, I. Bendib and M. Y. Haouam, "Breast cancer lesion detection and segmentation based on mask R-CNN," in *2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*, 2021.

[7] H. Yu, R. Bai, J. An and R. Cao, "Deep learning-based fully automated detection and segmentation of breast mass," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2020.

[8] D. Liu, B. Wu, C. Li, Z. Sun and N. Zhang, "TrEnD: A transformer-based encoder-decoder model with adaptive patch embedding for mass segmentation in mammograms," *Medical Physics,* vol. 50, p. 2884–2899, 2023.

[9] Y. Su, Q. Liu, W. Xie and P. Hu, "YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms," *Computer Methods and Programs in Biomedicine,* vol. 221, p. 106903, 2022.

[10] G. Jocher, "YOLOv5 by ultralytics," *Released date,* p. 5–29, 2020.

[11] C.-Y. Wang, A. Bochkovskiy and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[12] G. Jocher, A. Chaurasia and J. Qiu, "YOLO by Ultralytics," *URL: https://github. com/ultralytics/ultralytics,* 2023.

[13] D. Dluznevskij, P. Stefanovič and S. Ramanauskaite, "Investigation of YOLOv5 Efficiency in iPhone Supported Systems.," *Baltic Journal of Modern Computing,* vol. 9, 2021.

[14] S. Zhou, K. Cai, Y. Feng, X. Tang, H. Pang, J. He and X. Shi, "An Accurate Detection Model of Takifugu rubripes Using an Improved YOLO-V7 Network," *Journal of Marine Science and Engineering,* vol. 11, p. 1051, 2023.

[15] Ultralytics, "GitHub Issue 189 - Ultralytics," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics/issues/189.

[16] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data,* vol. 4, p. 1–9, 2017.

[17] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer Jr, R. Moore, K. Chang and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital Mammography: Nijmegen, 1998,* Springer, 1998, p. 457–460.

[18] L. Bouzar-Benlabiod, K. Harrar, L. Yamoun, M. Y. Khodja and M. A. Akhloufi, "A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification," *Computers in Biology and Medicine*, vol. 163, p. 107133, 2023.

[19] Y. Wang, S. Wang, J. Chen and C. Wu, "Whole mammographic mass segmentation using attention mechanism and multiscale pooling adversarial network," *Journal of Medical Imaging*, vol. 7, p. 054503–054503, 2020.

[20] G. Ayana, K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, N. Husen, F. Mulugeta, B. Habtamu and S.-w. Choe, "Vision-Transformer-Based Transfer Learning for Mammogram Classification," *Diagnostics*, vol. 13, p. 178, 2023.

[21] M. Cantone, C. Marrocco, F. Tortorella and A. Bria, "Convolutional Networks and Transformers for Mammography Classification: An Experimental Study," *Sensors*, vol. 23, p. 1229, 2023.

[22] H. Sun, C. Li, B. Liu, Z. Liu, M. Wang, H. Zheng, D. D. Feng and S. Wang, "AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms," *Physics in Medicine & Biology*, vol. 65, p. 055005, 2020.

[23] S. A. Kamran, K. F. Hossain, A. Tavakkoli, G. Bebis and S. Baker, "SWIN-SFTNet: Spatial Feature Expansion and Aggregation using Swin Transformer For Whole Breast micro-mass segmentation," *arXiv preprint arXiv:2211.08717*, 2022.